# Assignment-based Subjective Questions

# Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

# Total Marks: 3 marks (Do not edit)

# Answer: <Your answer for Question 1 goes below this line> (Do not edit)

# The categorical variables are season, yr, mnth, holiday, weekday, workingday, weathersit.

# 

# 

# As per the visualization done through box plot and bar plot:

# We can say that

# For season: Summer and fall has higher Median rental count.

# Winter has lower Median rental count.

# Spring has moderate Median rental count.

# For yr: 1 (2019) has higher demand

# 0 (2018) had lower demand

# For month: may to sep – higher rental

# Dec to feb – lower demand

# For weekday: working days has higher demand than weekends.

# For weathersit: Mist/cloudy – high demand

# Clear/few clouds – moderate demand

# Light snow/ light rain – low demand

# 

**Question 2.** Why is it important to use **drop\_first=True** during dummy variable creation? (Do not edit)

**Total Marks:** 2 marks (Do not edit)

# Answer: <Your answer for Question 2 goes below this line> (Do not edit)

# drop\_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation.it helps to avoid multicollinearity and ensures that the model coefficients are interpretable.

# Let’s say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not furnished and semi\_furnished, then It is obvious unfurnished. So we do not need 3rd variable to identify the unfurnished. Example

# Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables

**Question 3.** Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

**Total Marks:** 1 mark (Do not edit)

# Answer: <Your answer for Question 3 goes below this line> (Do not edit)

# The scatter plot depicting the relationship between 'temp' and 'atemp' reveals a strong linear correlation, indicating that these two variables are closely related. This suggests that 'atemp' can be reliably predicted from 'temp' with minimal error, making them nearly interchangeable in modeling scenarios.

**Question 4.** How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

**Total Marks:** 3 marks (Do not edit)

# Answer: <Your answer for Question 4 goes below this line> (Do not edit)

# 1. Residuals' Normality:

# To assess whether the residuals follow a normal distribution I calculated the differences between actual and predicted values and visualized them using a histogram.

# 2. Checking for Homoscedasticity:

# I examined the consistency of variance by plotting residuals against predicted values in a scatter plot looking for any patterns that might indicate heteroscedasticity.

# 3. Residual Independence:

# To identify any autocorrelation in the residuals I applied the Durbin-Watson test. A result close to 2 suggests that the residuals are not correlated with each other.

**Question 5.** Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)

**Total Marks:** 2 marks (Do not edit)

# Answer: <Your answer for Question 5 goes below this line> (Do not edit)

# Temp, yr and weather sit are the top 3 features contributing significantly towards the demand of the shared bikes

# General Subjective Questions

**Question 6.** Explain the linear regression algorithm in detail. (Do not edit)

**Total Marks:** 4 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

# <Your answer for Question 6 goes here>

# Linear regression is a statistical method used to predict a dependent variable (the target) based on one or more independent variables (the predictors). It works by finding the best-fitting straight line through the data points.

# For a single predictor, the formula is: Y = β₀ + β₁X + ε

# Where:

# Y is the target variable.

# X is the predictor variable.

# β₀ is the y-intercept (where the line crosses the Y-axis).

# β₁ is the slope (how much Y changes for a one-unit change in X).

# ε represents the error (the difference between the actual and predicted values).

# The process involves:

# Preparing the Data: Gathering and cleaning the data.

# Training the Model: Finding the line that best fits the data by minimizing the errors, typically using Ordinary Least Squares (OLS).

# Evaluating the Model: Checking how well the model performs using metrics like R-squared, Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).

# Making Predictions: Using the trained model to predict the target variable for new data.

# There are two main types:

# Simple Linear Regression: Uses one predictor.

# Multiple Linear Regression: Uses two or more predictors.

**Question 7.** Explain the Anscombe’s quartet in detail. (Do not edit)

**Total Marks:** 3 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

# <Your answer for Question 7 goes here>

# Anscombe's Quartet: Four datasets with nearly identical summary statistics (mean, variance, correlation, regression line) but drastically different visual patterns when graphed. Created by Francis Anscombe to emphasize the importance of visualizing data, it shows that relying solely on numerical summaries can be deceptive. Visualizing data (e.g., scatter plots) is essential for accurate interpretation and model building.

**Question 8.** What is Pearson’s R? (Do not edit)

**Total Marks:** 3 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

# <Your answer for Question 8 goes here>

# Pearson’s R measures the strength and direction of a linear relationship between two continuous variables.

# Key Points:

# Range: −1 to +1

# +1 -> Perfect positive correlation

# 0 -> No correlation

# −1 -> Perfect negative correlation

# Assumptions:

# Linear relationship

# No extreme outliers

# Continuous, normally distributed data

**Question 9.** What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)

**Total Marks:** 3 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

# <Your answer for Question 9 goes here>

# Scaling adjusts the range of data to ensure each feature contributes equally to the model's performance, especially in algorithms sensitive to feature scale.

# Scaling Performed is performed for:

# Equal contribution: Prevents features with larger ranges from dominating the model.

# Improves convergence: Helps algorithms that use gradient-based optimization converge faster.

# Reduces bias: Ensures features contribute equally to the model.

# Normalized vs. Standardized Scaling

# Normalized Scaling (Min-Max Scaling):

# Formula:

# Normalized Value = (x - min(X)) / (max(X) - min(X))

# Use Case: Best for when data needs to be in a specific range (e.g., neural networks).

# Standardized Scaling (Z-Score Scaling):

# Formula:

# Standardized Value = (x - μ) / σ

# where μ is the mean and σ is the standard deviation.

# Use Case: Useful when data is normally distributed or when models assume normality.

# Key Differences:

# Range: Normalized scaling compresses data to a fixed range (0 to 1), while standardized scaling gives data a mean of 0 and a standard deviation of 1.

# Use Case: Normalization is used when scale matters (e.g., distance-based algorithms), while standardization is for algorithms assuming normality.

**Question 10.** You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)

**Total Marks:** 3 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

# <Your answer for Question 10 goes here>

# VIF becomes infinite when there is perfect multicollinearity, meaning one predictor is a perfect linear combination of others. This leads to a singular correlation matrix, making the VIF calculation undefined. In such cases, you should remove or combine collinear predictors to improve model stability.

**Question 11.** What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

(Do not edit)

**Total Marks:** 3 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

# <Your answer for Question 11 goes here>

# A Q-Q plot compares the quantiles of a dataset to a theoretical distribution, typically the normal distribution. In linear regression, it is used to check if the residuals are normally distributed. If the points align along a straight line, it indicates normality. Deviations suggest problems like outliers, skewness, or nonlinearity, which may affect the model’s validity.